DOCUMENT RESUME

ED 464 937 TM 033 868

AUTHOR Huang, Chi-Yu; Lohss, William E.; Lin, Chuan-Ju; Shin, David
TITLE Item Calibrations for Licensure Tests with Multiple Specialty Components.
PUB DATE 2002-04-00
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Item Response Theory; *Licensing Examinations (Professions); *Test Items
IDENTIFIERS *BILOG Computer Program; *Calibration; PIC Program

ABSTRACT

This study was conducted to compare the usefulness of three item response theory (IRT) calibration packages (BILOG, BILOG-MG, and PIC) for examinations that include common and specialty components. Because small sample sizes and different mean abilities between specialty components are the most frequent problems that licensure/certification programs may encounter, this study emphasizes these issues. The three-parameter logistic model item calibrations for 360 items from a large-scale test were assumed to be the known item parameters and were used to generate data. Results show that BILOG, BILOG-MG, and PIC are equally successful in estimating item parameters when specialty groups have the same ability level However, BILOG-MG performs better than the other two packages when dealing with specialty groups with different ability levels. The study also suggests that the precision of item parameter estimates is not seriously affected when specialty groups have different ability levels. The specialty groups sample size has more effect on the precision of estimated item parameters when it reduces to 250. Although item characteristic curve plotting for some items in this study has shown that these estimation errors may be acceptable, the generalization of the results from this study to real testing situations should be made cautiously. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 937

C. Huang

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Item Calibrations for Licensure Tests with Multiple Specialty Components

Chi-Yu Huang, William E. Lohss, Chuan-Ju Lin

ACT, Inc.

David Shin

The University of Iowa

Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA

TM033868

BEST COPY AVAILABLE

Item Calibrations for Licensure Tests with Multiple Specialty Components

Chi-Yu Huang, William E. Lohss, Chuan-Ju Lin
ACT, Inc.

David Shin
The University of Iowa

Introduction

In response to increasing specialization within many occupations and professions, several licensure and certification programs now provide exams that not only evaluate general competency, but also the competencies required for various specialty areas. For example, an exam used for licensing engineers includes different components for several engineering specialties: chemical, civil, electrical, environmental, industrial, and mechanical engineering. All examinees are required to complete the common component along with one of the specialty components.

The management of item data, as well as the assembly of new exams, becomes a challenge for examination programs that include both common and specialty components. Should a single set of item statistics be computed over all examinees and stored as a single item bank or should different sets of item statistics be computed and banked for each specialty group of examinees? If multiple sets of item statistics are computed and banked, then how should the common component be assembled and equated? How can the item statistics and examinee ability across the different specialty exams be compared? From the perspective of simplicity of data management and test assembly, a single set of item statistics is preferred. This allows a single common component to be assembled for all examinees with one set of item statistics. This also makes possible the comparison of abilities between examinees who take different specialty exams.

There are three possible procedures for calibrating items for exams that incorporate specialty components using the Item Response Theory (IRT) approach. One procedure is to calibrate the common and specialty items separately for each group of examinees taking a different specialty exam. Doing this results in multiple sets of item calibrations for the common items, as well as calibrations for each set of specialty items that are based on a different group of examinees. A second procedure is to calibrate all

the common and specialty items together over all the examinees. This results in a single set of item calibrations for both the common and specialty items. The third procedure is to calibrate the common items using all examinees first. These calibrated item parameters are then fixed for estimating the specialty items. This procedure will also allow us to get a single set of item calibrations for the common and specialty items.

The present study focuses on the second and third procedures described above for exams that have both common and specialty components. These two procedures enable the simplification of data management and assembly of new exams, as well as the creation of an IRT item bank for future computerized exams. The IRT approach assumes that the same dominant latent ability is being measured across the sub-populations defined by the specialty exams. For most testing programs, BILOG is the most widely used 3-PL calibration software package. BILOG assumes that the latent abilities of all examinees are drawn from one normal distribution (Mislevy & Bock, 1990). However, in practice, real data sets may not satisfy this assumption. Some specialty components may require examinees to have different ability levels to pass. Therefore, examinees from different specialties may come from sub-populations that are normal but with different means and standard deviations. This situation could violate the assumption on which BILOG is based and affect the applicability of this program.

The BILOG-MG program, on the other hand, is an extension of IRT that is used for multiple groups of examinees. It uses information relating to group membership to improve the accuracy of item parameter calibration and ability estimation (Zimowski, Muraki, Mislevy & Bock, 1999). However, it becomes a concern when BILOG-MG is used for licensure/certification tests for which only small sample sizes are available for each specialty component. For some specialty areas that require higher ability, there may be only a small numbers of examinees. The effects of the interaction between ability differences and sample size on BILOG-MG item calibrations are not clear. In addition to the above two packages, a program called Pretest Item Calibration (i.e., PIC; Davey, 1990) has been successfully used to calibrate pretest items that were collected from small sample sizes, following Bayesian estimation procedures. When using the PIC program, items that all examinees answered are calibrated first. These item parameter estimates are used to compute the posterior ability distribution for all examinees. Item parameter estimates for items that are only answered by a portion of the examinees then are estimated individually with respect to the posterior. This program could be adapted to fit

the current situation. In other words, all of items in common component are calibrated together first, and based on the fixed common item parameters, items in specialty components are estimated individually. Although sample size may not be an issue when PIC is used, it is still unclear whether item calibration will be affected by ability differences of groups.

The purpose of this study is to compare the utility of the above IRT calibration packages (BILOG, BILOG-MG, and PIC) for exams that include common and specialty components. Because small sample sizes and different mean abilities between specialty components are the most frequent problems that licensure/certification programs may encounter, this study emphasizes the following issues when the results of the three calibration software packages are evaluated:

1. What is the effect of group sample size on the precision of item parameter estimates?
2. What is the effect on item parameter estimates when the specialty groups have different ability levels?
3. How are item parameter estimates affected by the interaction between sample size and ability difference across groups?

Procedure

The 3-PL item calibrations for 360 items from a large-scale test were assumed to be the known item parameters and were used to generate 0/1 data. The first 120 items were assumed to be the items making up the common component. The next 240 items were assumed to be four sets of 60 specialty items for four different specialty components (see Table 1).

Table 1: Items Included in Each Specialty Exam

	Exam A	Exam B	Exam C	Exam D
Item 1 to Item 120 (common items)	X	X	X	X
Item 121 to Item 180 (specialty A)	X			
Item 181 to Item 240 (specialty B)		X		
Item 241 to Item 300 (specialty C)			X	
Item 301 to Item 360 (specialty D)				X

Two factors that might affect item calibrations were examined: sample size and the various ability levels of examinees for the four specialty exam groups. Group sample sizes were selected to range between 250 and 1500. These values were selected because 1000 is the recommended minimum for BILOG; however, in many certification programs psychometricians must often attempt to calibrate items with smaller sample sizes, sometimes as small as 250.

The 0/1 response data were simulated using 3-PL model for 14 different cases that manipulated these two factors, as follows in Table 2:

Table 2: Sample Sizes and Ability Distributions of Simulation Groups

Case	Group Sample Size	θ Distribution	Case	Group Sample Size	θ Distribution
1	A=1500 B=1500 C=1500 D=1500	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	8	A=1500 B=1500 C=1500 D=1500	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
2	A=1500 B=1500 C=1000 D=500	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	9	A=1500 B=1500 C=1000 D=500	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
3	A=1500 B=1500 C=1000 D=250	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	10	A=1500 B=1500 C=1000 D=250	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
4	A=1500 B=1500 C=500 D=500	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	11	A=1500 B=1500 C=500 D=500	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
5	A=1500 B=1500 C=250 D=250	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	12	A=1500 B=1500 C=250 D=250	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
6	A=500 B=500 C=500 D=500	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	13	A=500 B=500 C=500 D=500	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)
7	A=250 B=250 C=250 D=250	θ~N(0,1) θ~N(0,1) θ~N(0,1) θ~N(0,1)	14	A=250 B=250 C=250 D=250	θ~N(0,1) θ~N(1,1) θ~N(0,1) θ~N(1,1)

For each of the 14 cases, the generated 0/1 data were used to calibrate item parameters for 360 items over all examinees using BILOG, BILOG-MG, and PIC.

Two criteria were used to evaluate the item parameter estimates for each case: (1) the root mean squared difference ($RMSD_\phi$), which can be calculated as

$$RMSD_\phi = \sqrt{\frac{1}{L}(\phi - \hat{\phi})^2},$$

where L = the number of items (60) for each specialty component

ϕ = the known a- or b- parameter;

and $\hat{\phi}$ = the estimated a- or b- parameter;

and (2) the comparison between known and estimated item characteristic curves (ICCs) for each item. For each item i, the *RMSD* between estimated and known ICCs could be weighted by the ability distribution, and be expressed as $delta_{ICC}$. The $delta_{ICC}$ can be calculated as

$$\Delta_{ICC_i} = \sqrt{\int [P_{est.}(U=1 \mid \hat{a}_i, \hat{b}_i, \hat{c}_i, \theta) - P_{known}(U=1 \mid a_i, b_i, c_i, \theta)]^2 \phi(\theta) d\theta},$$

where $\phi(\theta)$ is the normal density function that is used as the ability distribution in the study. The mean of Δ_{ICC} for each specialty component ($\overline{\Delta}_{ICC}$) is the index of ICC recovery (Hsu, Thompson, & Chen, 1998). In other words,

$$\overline{\Delta}_{ICC} = \frac{1}{L}\sum_{i=1}^{L} \Delta_{ICC_i},$$

where L is the number of items (60) in each specialty component.

Results and Discussion

Table 3 lists the *RMSD* values for the a- and b- parameters from BILOG results. The *RMSD* values between known and estimated parameters from BILOG-MG and PIC are listed in Table 4 and Table 5. As shown in Table 3 through Table 5, when subgroups had the same ability distribution (Case1 to Case7), the *RMSD* values for the a- and b- parameters resulting from each software program were similar. However, when subgroups had different ability levels (Case8 to Case14), the overall a-parameter estimates from BILOG had higher *RMSD* values than PIC and BILOG-MG, and the overall b-parameter estimates from BILOG-MG had smallest *RMSD* values. It appears

that the performance of BILOG-MG is very similar to that of BILOG and PIC when subgroups come from the same ability level, and it is superior to BILOG and PIC when subgroups have different ability levels.

Table 6 lists the $\overline{\Delta}_{ICC}$ values for each common item and specialty component from three programs. Again, as shown for Case1 to Case7, the $\overline{\Delta}_{ICC}$ values in each component from BILOG-MG were very similar to those from BILOG and PIC when subgroups came from the same ability level. The $\overline{\Delta}_{ICC}$ values from BILOG-MG were much smaller than those from BILOG and PIC when subgroups had different ability levels, which could be observed from results of Case8 through Case14. It implies, regardless of sample size for each specialty group, that BILOG and PIC could not calibrate item parameters as precisely as BILOG-MG when groups were involved various ability levels. Because the results from BILOG-MG were similar or superior to the other two programs, the following discussion will be based only on the results from BILOG-MG.

Table 4 shows, as expected, that when the sample size of the specialty groups decreased, the *RMSD* values gradually increased for both the *a*- and *b*- parameters of items in the common and specialty components. In other words, estimation error became greater. However, the results from subgroups having the same ability level (Case1 to Case7) and from subgroups with different ability levels (Case8 to Case14) had similar patterns. This indicates that differences in ability levels do not affect the precision of concurrent calibration for common and specialty items. The sample size has more effect on the precision of item parameter estimates than same/different ability levels.

In addition, based on the comparison of known and estimated ICCs for items, the index of ICC recovery ($\overline{\Delta}_{ICC}$) provides a more heuristic evaluation because ICCs may be relatively unaffected by large errors in the estimation of item parameters (Hulin, Drasgow, & Parsons, 1983). The $\overline{\Delta}_{ICC}$ values for the BILOG-MG are shown in Table 6. Similar to the *RMSD* results, when the sample sizes of specialty groups decreased, the $\overline{\Delta}_{ICC}$ values gradually increased in common and specialty components. This increase was more obvious when the sample size of specialty group was 250. Again, the results from subgroups having the same ability level (Case1 to Case7) and from subgroups having different ability levels (Case8 to Case14) were very similar.

To further investigate how different the ICCs from various sample sizes are, several items, which have bigger values of Δ_{ICC} in Case14, were chosen to compare the plots of ICCs from the different conditions. The ICC from Case1 served as the base because it had most similar curve to the known ICC. The 95% confidence bands for Case1's ICC were also plotted. Figure 1 and Figure 2 show the curves for item 14 from specialty B, and item 8 from specialty D.

The plots for item 14 of specialty B shown in Figure 1 compare the ICCs from different sample sizes and ability levels with the base ICC from the Case1 (N = 1500 in each subgroup with the same ability). As can be seen from plots, although most ICCs were not identical to Case1's ICC, they still fell within the 95% confidence bands. Only part of the ICC from Case14 (small sample size with various ability levels) was out of Case1's 95% confidence bands in the lower ability ranges. Plots in Figure 2 also present similar patterns. Most of the cases' ICCs fell within the 95% confidence bands. This suggests that, even though $\overline{\Delta}_{ICC}$ from different cases have some discrepancy, when compared with the estimation from data with the big sample size and having the same ability level, these errors could be tolerated. The specialty groups' sample sizes and ability levels do not seriously affect the precision of item parameter estimates. For these two items, ICCs outside of the 95% confidence bands were both shown in the lower ability range. This phenomenon may happen because there were even fewer examinees in the end of lower ability tail when the sample size became smaller. As a result, greater estimation errors could be expected in the range of lower ability. For most licensure and certification programs, the purpose of the testing is to classify examinees as to whether they are pass or fail, but not to measure their true ability; and the decision point (or cutting point) is usually set in the middle range of the ability. Because of this, it might possible to ignore the estimation error in the lower ability ranges.

Conclusion

When dealing with IRT-based examinations that include both common and specialty components, there are several advantages to be gained by calibrating items over all the examinees vs. producing several sets of different calibrations based on each specialty exam group. This study has shown that, BILOG, BILOG-MG and PIC are equally successful in estimating item parameters when specialty groups have the same

ability level. However, BILOG-MG performs better than the other two packages when dealing with specialty groups with different ability levels. This study also suggests that the precision of item parameter estimates is not seriously affected when specialty groups have different ability levels. The specialty groups' sample size has more effect on the precision of estimated item parameters when it reduces to 250. Although ICC plotting for some items in this study has shown that these estimation errors may be acceptable, the generalization of the results from this study to real testing situation should be made cautiously.

References

Davey, T. (1990). *Pretest item calibration*. Unpublished manuscript.

Hsu, Y., Thompson, T.D., & Chen, W.-H. (1998, April). *CAT Item calibration*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item response theory: Application to psychological measurement. Homewook, IL: Dow Jones-Irwin.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Chicago: Scientific Software International.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1999). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago: Scientific Software International.

Table3: The RMSD Values for *a*- and *b*- Parameters from Bilog Results

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13	Case 14
***a* Parameter**														
Common	0.0634	0.0729	0.0735	0.0656	0.0786	0.0931	0.1211	0.1280	0.1440	0.1302	0.1376	0.1138	0.1307	0.1578
Specialty A	0.1488	0.1555	0.1541	0.1413	0.1418	0.1456	0.2288	0.1970	0.1530	0.1563	0.1870	0.1746	0.2147	0.2548
Specialty B	0.1115	0.0902	0.1056	0.0926	0.1150	0.1518	0.1755	0.1570	0.1500	0.1591	0.1665	0.1470	0.2046	0.2106
Specialty C	0.1273	0.1326	0.1508	0.1471	0.3173	0.2481	0.2666	0.1910	0.2110	0.1805	0.1963	0.3246	0.2244	0.2333
Specialty D	0.1557	0.2215	0.2580	0.2250	0.2352	0.2752	0.2626	0.1630	0.2270	0.2306	0.2162	0.2252	0.2027	0.2549
***b* Parameter**														
Common	0.1525	0.1254	0.1746	0.1407	0.1221	0.1461	0.1881	0.5700	0.4650	0.4619	0.5483	0.5522	0.5867	0.5375
Specialty A	0.1813	0.1682	0.1673	0.1462	0.1883	0.2103	0.3051	0.5520	0.4670	0.4689	0.4995	0.5290	0.6026	0.5957
Specialty B	0.1587	0.1698	0.1628	0.1340	0.1409	0.2362	0.3073	0.5510	0.4570	0.4450	0.5186	0.5251	0.5588	0.5644
Specialty C	0.1707	0.1467	0.1892	0.2751	0.3323	0.2133	0.3205	0.5420	0.4170	0.4386	0.5092	0.5819	0.5259	0.5455
Specialty D	0.1438	0.1900	0.3377	0.2449	0.2501	0.2373	0.2479	0.5460	0.5080	0.5221	0.5703	0.5557	0.5929	0.5190

Table 4: The RMSD Values for *a*- and *b*- Parameters from Bilog-MG Results.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13	Case 14
***a* Parameter**														
Common	0.0668	0.0740	0.0713	0.0740	0.0780	0.1231	0.1248	0.0570	0.0739	0.0654	0.0706	0.0700	0.0795	0.1044
Specialty A	0.1369	0.1541	0.1555	0.1427	0.1387	0.1393	0.2281	0.1387	0.1374	0.1379	0.1413	0.1457	0.1783	0.2031
Specialty B	0.1180	0.0977	0.1061	0.0963	0.1138	0.1679	0.1659	0.0929	0.0982	0.1044	0.1006	0.1101	0.1551	0.1541
Specialty C	0.1368	0.1373	0.1507	0.1599	0.2976	0.2409	0.2615	0.1226	0.1512	0.1448	0.1747	0.2960	0.1989	0.2361
Specialty D	0.1547	0.2167	0.2568	0.1950	0.2602	0.2866	0.2943	0.1002	0.1896	0.2071	0.1897	0.2203	0.1690	0.2334
***b* Parameter**														
Common	0.1575	0.1196	0.1766	0.1465	0.1255	0.1669	0.2178	0.1393	0.1358	0.1161	0.1446	0.1438	0.1747	0.2224
Specialty A	0.1742	0.1633	0.1665	0.1374	0.1898	0.2102	0.3206	0.1106	0.1663	0.1771	0.1706	0.1407	0.1925	0.2401
Specialty B	0.1549	0.1661	0.1688	0.1268	0.1453	0.2351	0.3505	0.1612	0.1684	0.1796	0.1530	0.1741	0.2073	0.3164
Specialty C	0.1735	0.1328	0.1929	0.2843	0.3149	0.2117	0.3417	0.1782	0.1550	0.1936	0.2282	0.2921	0.1990	0.3315
Specialty D	0.1428	0.1825	0.3402	0.2466	0.2654	0.2309	0.3034	0.1574	0.2122	0.2735	0.1854	0.2679	0.2495	0.2651

Table 5: The RMSD Values for *a*- and *b*- Parameters from PIC Results.

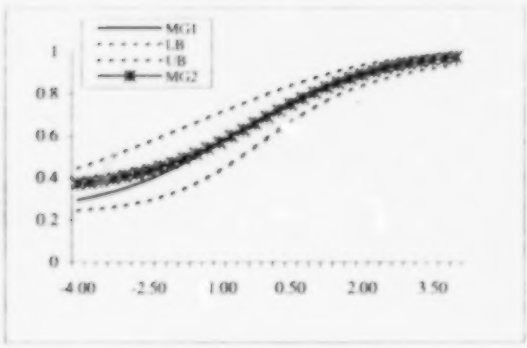
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13	Case 14
***a* Parameter**														
Common	0.0636	0.0734	0.0758	0.0680	0.0828	0.0943	0.1274	0.1273	0.1437	0.1313	0.1271	0.1098	0.1292	0.1530
Specialty A	0.0989	0.1056	0.3141	0.1253	0.0949	0.2340	0.3214	0.1607	0.1030	0.0860	0.1596	0.0995	0.1971	0.2702
Specialty B	0.0912	0.0834	0.1539	0.0715	0.0818	0.1689	0.1890	0.0924	0.0986	0.0994	0.0834	0.0803	0.1416	0.1830
Specialty C	0.0989	0.1327	0.1861	0.1825	0.3131	0.2787	0.2772	0.0866	0.1420	0.1086	0.1667	0.3133	0.2246	0.2164
Specialty D	0.0782	0.2213	0.3105	0.2288	0.2741	0.2129	0.2982	0.1077	0.1534	0.2099	0.2021	0.2053	0.1465	0.2590
***b* Parameter**														
Common	0.1506	0.1217	0.1612	0.1356	0.1240	0.1479	0.1813	0.5389	0.4341	0.4396	0.5165	0.5314	0.5487	0.5013
Specialty A	0.2100	0.1794	0.1960	0.2150	0.1857	0.2227	0.3131	0.4171	0.3349	0.3365	0.4226	0.4595	0.5862	0.6160
Specialty B	0.2543	0.2101	0.2195	0.1873	0.2079	0.2557	0.2843	0.5613	0.5041	0.4794	0.5520	0.5869	0.6166	0.6477
Specialty C	0.2815	0.2252	0.2429	0.2678	0.2860	0.2286	0.2928	0.4672	0.3721	0.3954	0.5204	0.6115	0.5291	0.5798
Specialty D	0.1877	0.2285	0.3049	0.2659	0.2637	0.2505	0.2333	0.5693	0.6123	0.6112	0.6554	0.6436	0.6532	0.6155

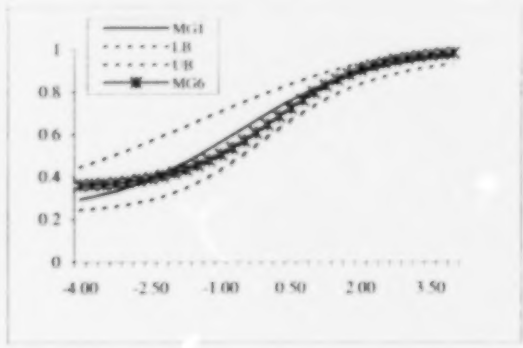
Table 6: Comparison of ICC Recovery for Each Program in Different Cases

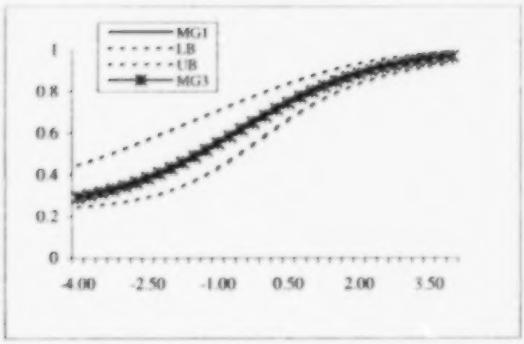
	Case1			Case2			Case3			Case4			Case5			Case6			Case7		
	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC
Common																					
mean	0.0094	0.0096	0.0094	0.0112	0.0111	0.0113	0.0138	0.0133	0.0139	0.0106	0.0113	0.0109	0.0110	0.0114	0.0111	0.0152	0.0190	0.0153	0.0195	0.0230	0.0197
SD	0.0042	0.0042	0.0043	0.0051	0.0049	0.0052	0.0067	0.0065	0.0067	0.0050	0.0048	0.0048	0.0051	0.0048	0.0056	0.0066	0.0074	0.0066	0.0092	0.0103	0.0095
A																					
mean	0.0186	0.0183	0.0197	0.0203	0.0200	0.0217	0.0191	0.0189	0.0402	0.0170	0.0167	0.0225	0.0176	0.0178	0.0223	0.0239	0.0256	0.0301	0.0394	0.0392	0.0417
SD	0.0081	0.0081	0.0099	0.0096	0.0078	0.0122	0.0072	0.0072	0.0532	0.0091	0.0087	0.0114	0.0085	0.0083	0.0097	0.0123	0.0137	0.0124	0.0199	0.0203	0.0210
B																					
mean	0.0161	0.0165	0.0167	0.0180	0.0189	0.0195	0.0193	0.0192	0.0228	0.0150	0.0156	0.0182	0.0174	0.0177	0.0172	0.0279	0.0286	0.0315	0.0408	0.0427	0.0403
SD	0.0077	0.0074	0.0066	0.0076	0.0081	0.0092	0.0084	0.0082	0.0100	0.0071	0.0070	0.0096	0.0071	0.0069	0.0063	0.0125	0.0138	0.0134	0.0165	0.0180	0.0160
C																					
mean	0.0186	0.0187	0.0180	0.0208	0.0202	0.0238	0.0208	0.0207	0.0249	0.0274	0.0283	0.0307	0.0397	0.0399	0.0392	0.0298	0.0315	0.0330	0.0360	0.0385	0.0378
SD	0.0079	0.0076	0.0095	0.0082	0.0083	0.0096	0.0099	0.0097	0.0120	0.0120	0.0125	0.0133	0.0203	0.0208	0.0198	0.0124	0.0140	0.0129	0.0181	0.0198	0.0177
D																					
mean	0.0173	0.0173	0.0169	0.0280	0.0275	0.0293	0.0402	0.0403	0.0421	0.0292	0.0295	0.0304	0.0339	0.0337	0.0342	0.0295	0.0316	0.0300	0.0326	0.0340	0.0341
SD	0.0087	0.0083	0.0089	0.0112	0.0113	0.0115	0.0231	0.0230	0.0216	0.0138	0.0136	0.0149	0.0165	0.0170	0.0145	0.0140	0.0137	0.0151	0.0158	0.0156	0.0153

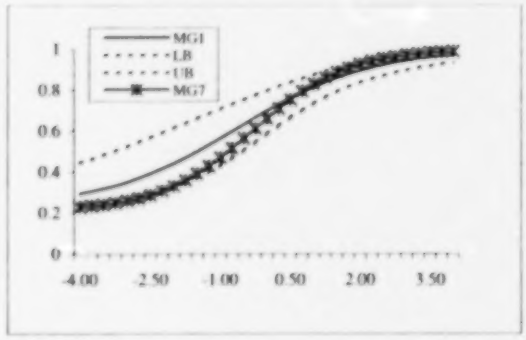
	Case8			Case9			Case10			Case11			Case12			Case13			Case14		
	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC	Bilog	MG	PIC
Common																					
mean	0.0984	0.0095	0.0915	0.0821	0.0105	0.0760	0.0807	0.0102	0.0754	0.0964	0.0108	0.0898	0.0946	0.0120	0.0895	0.0974	0.0163	0.0896	0.0946	0.0239	0.0882
SD	0.0293	0.0046	0.0273	0.0255	0.0046	0.0239	0.0242	0.0047	0.0226	0.0305	0.0052	0.0286	0.0297	0.0054	0.0284	0.0292	0.0073	0.0269	0.0283	0.0102	0.0265
A																					
mean	0.1013	0.0155	0.0735	0.0819	0.0183	0.0533	0.0803	0.0181	0.0550	0.0920	0.0175	0.0675	0.0963	0.0158	0.0756	0.1033	0.0260	0.0853	0.1024	0.0335	0.0870
SD	0.0300	0.0070	0.0225	0.0261	0.0091	0.0210	0.0246	0.0079	0.0195	0.0298	0.0077	0.0255	0.0249	0.0070	0.0209	0.0349	0.0135	0.0317	0.0343	0.0140	0.0344
B																					
mean	0.1036	0.0211	0.0950	0.0881	0.0208	0.0799	0.0852	0.0205	0.0781	0.1018	0.0197	0.0923	0.0998	0.0178	0.0953	0.1004	0.0298	0.0911	0.1018	0.0426	0.0956
SD	0.0361	0.0119	0.0360	0.0305	0.0118	0.0302	0.0267	0.0102	0.0265	0.0316	0.0097	0.0303	0.0324	0.0087	0.0315	0.0369	0.0155	0.0356	0.0394	0.0238	0.0382
C																					
mean	0.1057	0.0176	0.0671	0.0838	0.0207	0.0550	0.0871	0.0207	0.0584	0.0987	0.0276	0.0802	0.1115	0.0425	0.0948	0.1023	0.0268	0.0839	0.1008	0.0404	0.0891
SD	0.0374	0.0070	0.0182	0.0287	0.0087	0.0200	0.0320	0.0096	0.0230	0.0340	0.0121	0.0320	0.0461	0.0193	0.0411	0.0342	0.0114	0.0313	0.0378	0.0195	0.0366
D																					
mean	0.1101	0.0204	0.0968	0.0983	0.0313	0.0879	0.0982	0.0374	0.0913	0.1116	0.0291	0.1029	0.1111	0.0353	0.1038	0.1078	0.0334	0.0963	0.1018	0.0454	0.0939
SD	0.0368	0.0101	0.0324	0.0362	0.0168	0.0312	0.0398	0.0226	0.0378	0.0415	0.0185	0.0371	0.0439	0.0241	0.0415	0.0408	0.0161	0.0369	0.0455	0.0235	0.0411

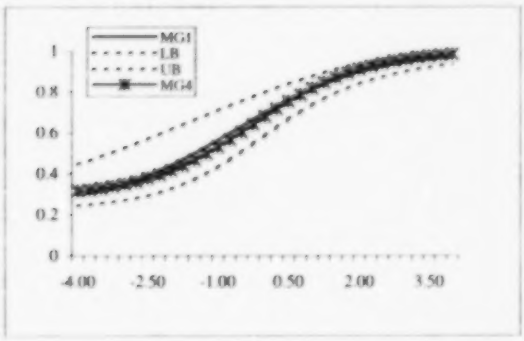
Figure1: Comparison of ICCs for Item 14 of Specialty B in Each Case

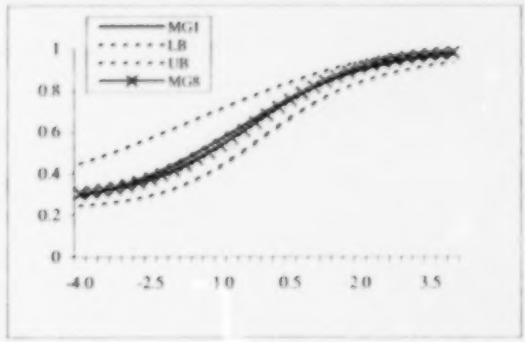


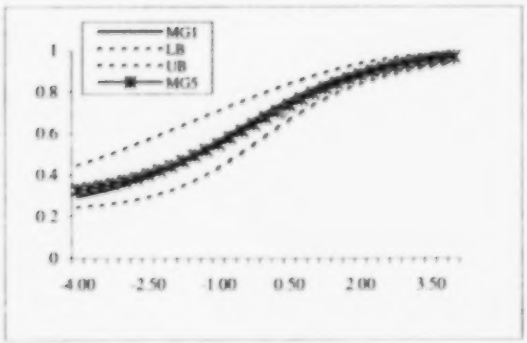


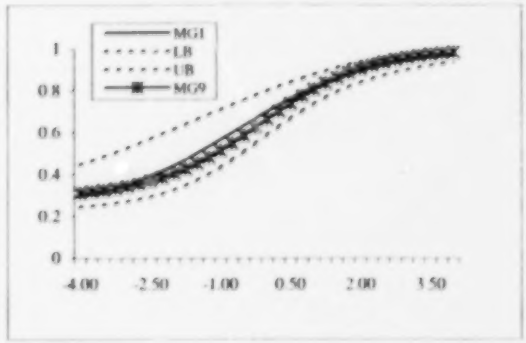






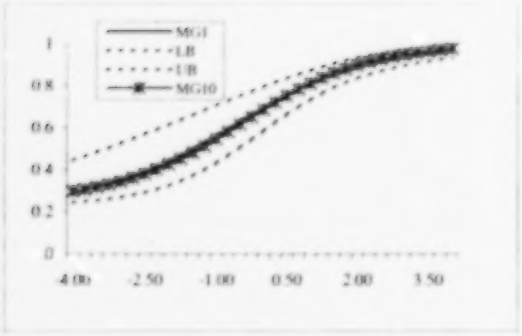


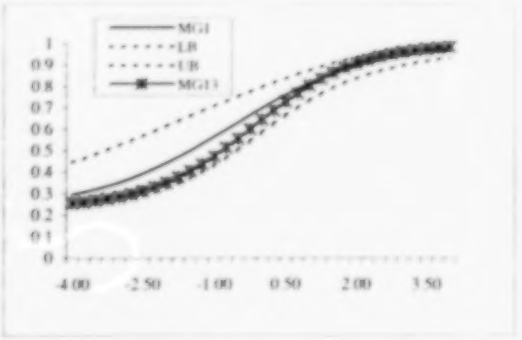


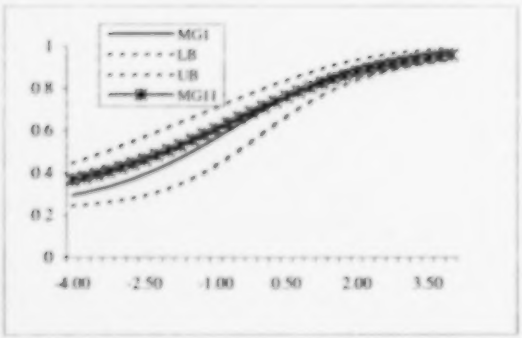


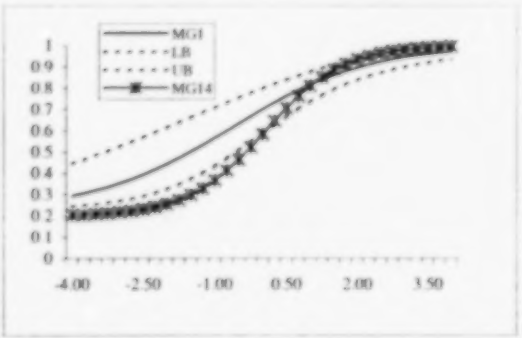
Note: MGX = ICC from CaseX's BILOG-MG result
LB = Lower band
UB = Upper band

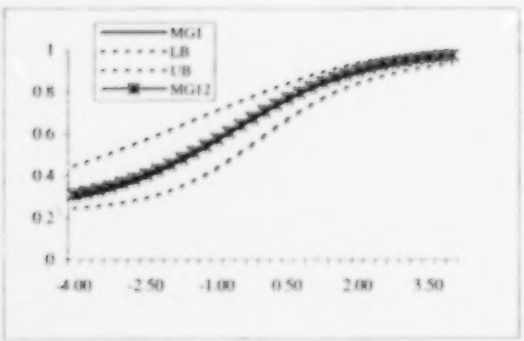
Figure1: Comparison of ICCs for Item 14 of Specialty B in Each Case





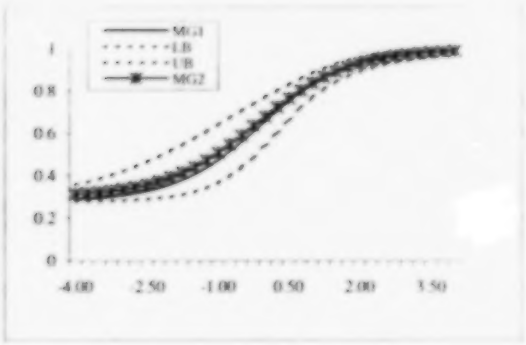


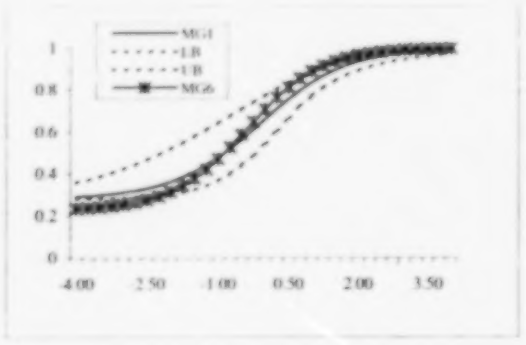


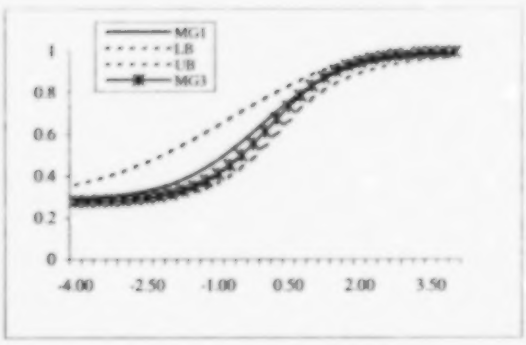


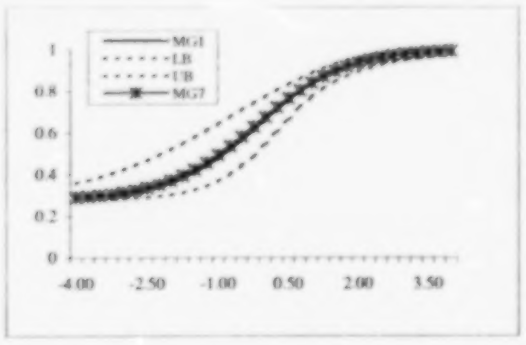
Note: MGX = ICC from CaseX's BILOG-MG result
LB = Lower band
UB = Upper band

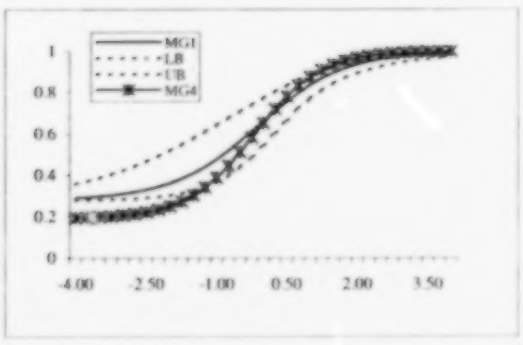
Figure 2: Comparison of ICCs for Item 8 of Specialty D in Each Case

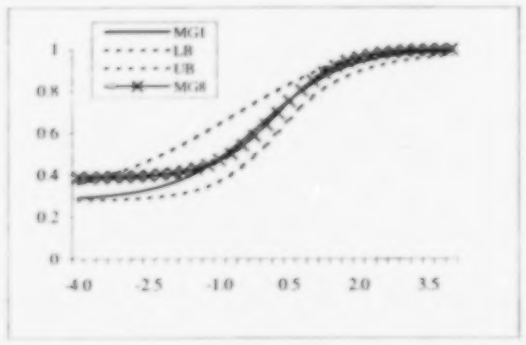


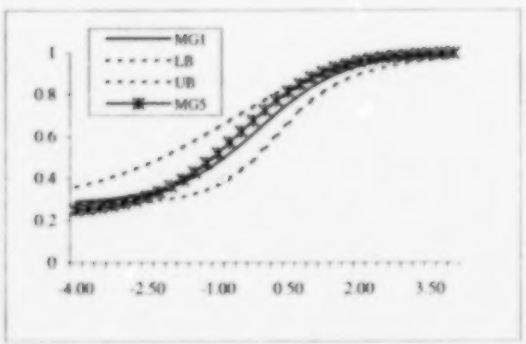


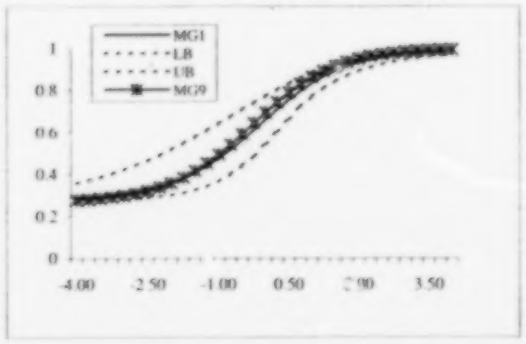






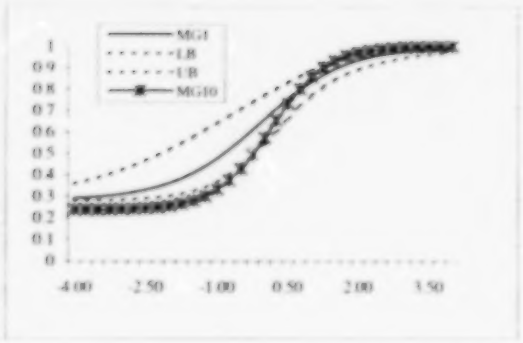


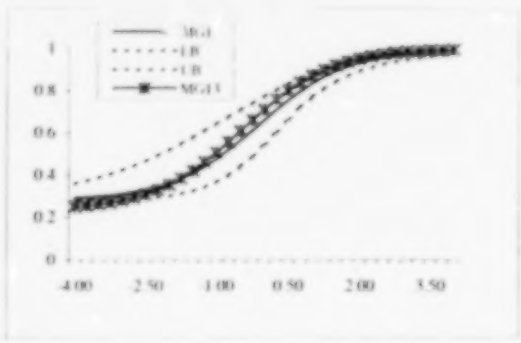


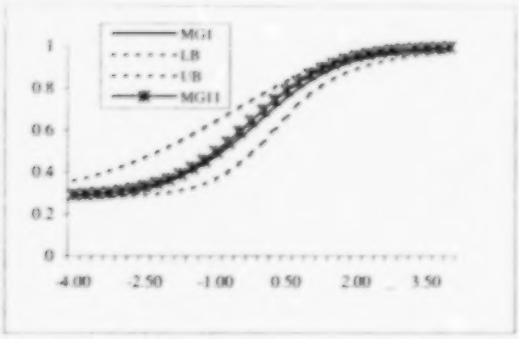


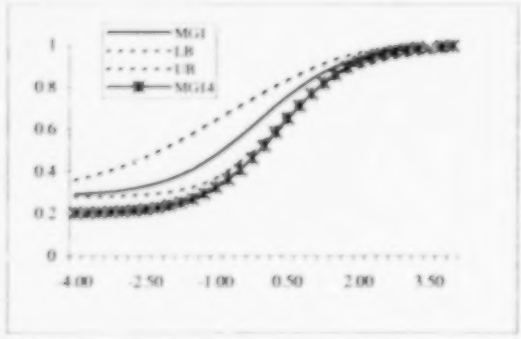
Note: MGX = ICC from CaseX's BILOG-MG result
LB = Lower band
UB = Upper band

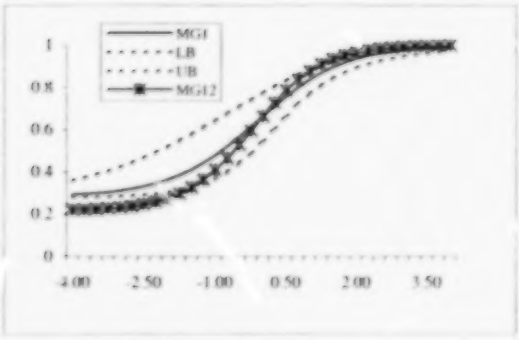
Figure 2: Comparison of ICCs for Item 8 of Specialty D in Each Case











Note: MGX = ICC from CaseX's BILOG-MG result
LB = Lower band
UB = Upper band